Chemometrics and Intelligent Laboratory Systems xxx (2013) xxx–xxx

Contents lists available at ScienceDirect

Chemometrics and Intelligent Laboratory Systems

journal homepage: www.elsevier.com/locate/chemolab

Identifying common and distinctive processes underlying multiset data

K. Van Deun [a,*], A.K. Smilde [b], L. Thorrez [c,d], H.A.L. Kiers [e], I. Van Mechelen [a]

[a] Research Group of Quantitative Psychology and Individual Differences, Faculty of Psychology and Educational Sciences, KU Leuven, Belgium
[b] Biosystems Data Analysis, Swammerdam Institute for Life Sciences, University of Amsterdam, Amsterdam, The Netherlands
[c] Interdisciplinary Research Facility Life Sciences, KU Leuven Kulak, Belgium
[d] Translational Cardiomyology, Department of Development and Regeneration, KU Leuven, Leuven, Belgium
[e] Heymans Institute, University of Groningen, Groningen, The Netherlands

ARTICLE INFO

Article history:
Received 31 December 2012
Received in revised form 8 July 2013
Accepted 12 July 2013
Available online xxxx

Keywords:
Multiset data
Common and distinctive
Data integration

ABSTRACT

In many research domains it has become a common practice to rely on multiple sources of data to study the same object of interest. Examples include a systems biology approach to immunology with collection of both gene expression data and immunological readouts for the same set of subjects, and the use of several high-throughput techniques for the same set of fermentation batches. A major challenge is to find the processes underlying such multiset data and to disentangle therein the common processes from those that are distinctive for a specific source. Several integrative methods have been proposed to address this challenge including canonical correlation analysis, simultaneous component analysis, OnPLS, generalized singular value decomposition, DISCO-SCA, and ECO-POWER. To get a better understanding 1) of the methods with respect to finding common and distinctive components and 2) of the relations between these methods, this paper brings the methods together and compares them both on a theoretical level and in terms of analyses of high-dimensional micro-array gene expression data obtained from subjects vaccinated against influenza.

1. Introduction

Technological developments have greatly facilitated obtaining, storing, and sharing data. This has created unique opportunities to study multiple aspects of a particular object of interest as witnessed by the recent surge of multifaceted approaches to several phenomena. For example, in systems vaccinology, the different components of the immune system are studied by means of, for example, genomewide measurements of mRNA transcription rates, cytokine and chemokine concentrations, and antibody responses, all measured on the same group of subjects [1]. Such data, consisting of multiple data blocks that are linked by the same set of units, are called multiset or multiblock data. Here, we focus on multiset data consisting of multiple object by variable data matrices that can be linked either by the objects or by the variables; this is one of the two orders is common, see [2] for a thorough discussion on the possible connections between multiple data matrices in the context of multivariate curve resolution. An example of objectwise linked data is the systems vaccinology example with several data matrices obtained for the same set of subjects. Another example from analytical chemistry is gas and liquid chromatography mass spectrometry data obtained for the same set of *Escherichia coli* fermentation batches [3]. An example of variable-wise linked data is gene expression data obtained for the same set of orthologous genes measured in different species [4].

Multiset data create a particular challenge for data analysis: often, they are not directly comparable (for example, due to differences in measurement scale between the different data matrices) while the aim is to find the mechanisms or processes that underlie all data blocks simultaneously. Interspecies comparative genomics, for example, aims at finding evolutionary conserved biological processes that are shared between all species-specific data matrices. Another example is the application of multiset multivariate curve resolution to different chromatographic runs in order to find the chemical components underlying all data blocks simultaneously [5]. To gain a deeper understanding of the processes underlying the data, it is often of primary importance to disentangle common and distinctive processes where common processes are processes that take place in all data blocks and distinctive processes are processes that take place in one data block or a few data blocks only. For example, in a systems approach to vaccinology an understanding of the immune system studied requires knowing which processes drive all parts of the immune system and which processes drive only specific parts of it. In the comparative genomics example, it is not only of importance to find the conserved (common) processes, but also the diverged (species-specific) ones. In multivariate curve resolution it may be of interest to find on the one hand common species with the same concentration profiles in all data blocks and on the other hand common species that have zero-concentration profiles in a particular block [2].

* Corresponding author at: Tiensestraat 102, 3000 Leuven, Belgium. Tel.: +32 16 32 58 88.
E-mail address: katrijn.vandeun@ppw.kuleuven.be (K. Van Deun).

0169-7439/$ – see front matter © 2013 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.chemolab.2013.07.005

To find common and sometimes also distinctive processes in multiset data, different dimension reduction methods have been proposed. The most promising dimension reduction methods are those that model all data blocks simultaneously because such a strategy avoids bias towards one particular data block (unlike strategies that first model one particular data block and subsequently use the derived model structure to represent the remaining data blocks, see for example [6]). Focusing on methods that treat the different data blocks in an exchangeable way (unlike regression methods) and that are exploratory in nature, several such methods have been proposed. To our knowledge, these are several variants of (generalized) canonical correlation analysis [7] and simultaneous component analysis [8,9], ECO-POWER [10], OnPLS [11], generalized singular value decomposition [12], and DISCO-SCA [13,14]. The first three methods deal with common components while the last three methods deal with both common and distinctive components. At present a discussion on how the methods deal with the issue of common and specific components has not yet been given and the relations between these methods are not yet well understood. To remedy for this, a first aim of the present paper is to bring these methods together and to compare them in terms of their formal description on a theoretical level, as well as in terms of analyses of a specific empirical data set. The latter will allow us to trace differences in finding common and distinctive processes at the practical level of data analysis. A challenging empirical case has been chosen, namely, micro-array gene expression data, which will allow us to evaluate the methods in the high-dimensional context and to interpret the output of the analyses by using functional annotation tools. Note that this paper differs considerably from our previous paper on simultaneous component analysis [9] as the latter does not discuss the issue of common and distinctive sources of variation and, on the level of the methods, does not include the adapted GSVD, DISCO-SCA, ECO-POWER, canonical correlation analysis, and OnPLS. It also differs from [14] that only compares DISCO-SCA, the GSVD, and the adapted GSVD.

The paper is organized as follows: first the illustrative data set will be introduced, then the different methods for finding common and specific processes will be briefly presented and applied to the illustrative data, followed by a thorough discussion of the different methods and of key issues in the search for common and specific processes. To support reproducibility all self-developed code used for reading, pre-processing, and analyzing the (publicly available) data and for writing output is provided at the ResearchGate page of the first author.

2. Data set

In a systems biology approach to vaccination against influenza, Nakaya et al. [1] studied simultaneously antibody response, cytokine concentrations, and the genome wide expression in humans vaccinated against influenza. Here, we concentrate on gene expression data for the 2008–2009 season (the series GSE29615 and GSE29617 publicly available at http://0-www.ncbi.nlm.nih.gov.ilsprod.lib.neu.edu/geo/; RMA was used for pre-processing the CEL files [15]). This is the only season with data for two different types of vaccines, namely trivalent inactivated influenza vaccine (TIV; complete data for 24 vaccinated young adults) or live attenuated influenza vaccine (LAIV; complete data for 27 vaccinated young adults). For each of the participants, a micro-array analysis was performed on the genomewide expression in peripheral blood mononuclear cells collected at baseline, and both 3 and 7 days after vaccination. The microarray used to collect the gene expression data includes 54,715 probe sets including approximately 38,500 well-characterized genes. We will include all probe sets in the analyses and the data at Day 3 and Day 7 are corrected for the baseline at Day 0 by taking the difference scores (for example, the Day 3 corrected data are obtained by subtracting the Day 0 data from the Day 3 data).

To illustrate the analysis of object- as well as variable-wise linked data, we will use these data in two ways. First, objectwise linked data will be created by combining the two corrected expression matrices of the 24 subjects vaccinated with TIV, one pertaining to the measurements at three and one to those at seven days after vaccination (see Fig. 1(a) for a graphical representation). The same set of 54,715 probe sets will be considered for each data block but we will treat these as two different sets of variables. Second, variable-wise linked data will be created by combining the baseline corrected expression matrix of the 24 TIV vaccinees at Day 3 with the baseline corrected expression matrix of the 27 LAIV vaccinees at Day 3 for the same set of 54,715 probe sets (see Fig. 1(b) for a graphical representation). To give equal weight to the variables, they will be centered and scaled to sum-of-squares one within each data block. Furthermore, each data block will be scaled to equal sum of squares such that the blocks have an equal weight in the analyses. Note that centering per block removes differences in means between the block; the focus is on the covariation between the genes (probe sets) and, more in particular, on the similarities and differences between the blocks in this intra-block covariance structure. There is systematic variation in vaccine efficacy between

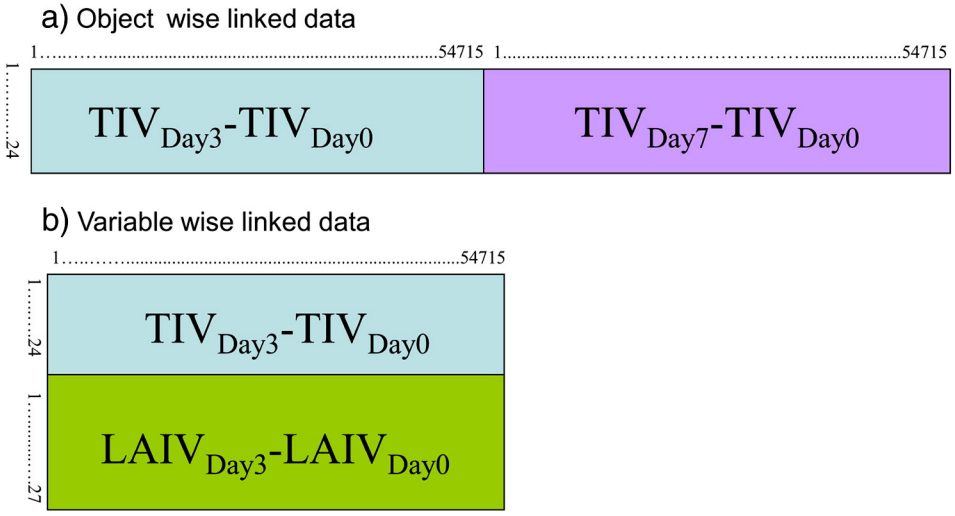

Fig. 1. Structure of the illustrative data. (a) Top panel: Objectwise linked data consisting of two data matrices with baseline corrected mRNA expression for samples collected three and seven days after vaccination in the same group of 24 vaccinated subjects. (b) Bottom panel: Variable-wise linked data consisting of two data matrices with the baseline corrected mRNA abundance at Day 3 for 54,715 probe sets obtained from two groups of subjects, 24 vaccinated with TIV and 27 with LAIV.

Please cite this article as: K. Van Deun, et al., Identifying common and distinctive processes underlying multiset data, Chemometrics and Intelligent Laboratory Systems (2013), http://dx.doi.org/10.1016/j.chemolab.2013.07.005

subjects (e.g., due to prior infections) that directly influences the innate and adaptive immune responses: therefore analysis of the covariation between transcripts may be expected to reveal immunological mechanisms by which the vaccine works. These mechanisms of action underlying a vaccine may be expected to be rather similar over time but maybe not between vaccines [1].

We will apply the different methods to the centered and scaled data and assess their biological relevance in two ways. First, we will use the results to predict the antibody response (the plasma hemagglutination–inhibition antibody titers) 28 days after vaccination. The antibody response of each vaccinee is included in the publicly available data (the series matrices GSE29615 and GSE29617) and is a measure of vaccine efficacy (with higher values indicating more efficacy). Second, to have an indication of the biological content, we will annotate the results by a publicly available annotation tool, Gene Set Enrichment Analysis (GSEA); see [16].

The input to the GSEA tool consists of pre-defined sets of genes with a common biological function and a ranked list of the genes. The pre-defined sets are based on common knowledge about gene function while the ranked list is based on the gene expression data at hand. At the moment of writing, a database specific for immunology related gene sets was released; we will rely on this database for the annotation of the results. For the given gene sets and the ranked list, GSEA tests for each gene set whether the genes belonging to the set are distributed randomly throughout the ranked list against the hypothesis that they are primarily located at the top or the bottom in the ranked list. The test statistic used is the enrichment score and its significance is assessed by relying on a permutation test procedure. To control for multiple testing, the family wise error rate is also provided. We refer to the original publication for more details [16].

3. Methods to identify common and specific underlying mechanisms

We will rely on the notation introduced by [17]: matrices are written in bold uppercase, vectors in bold lower case, and scalars in italic. The superscript $^{\mathrm{T}}$ is used to denote the transpose of matrices and vectors. The maximum value of a running index is denoted by the capital of the index letter. For example, the kth data block is denoted by $\mathbf{X}_k$ with k running from 1 to K. In this paper k is restricted to $k = 1{,}2$ (thus $K = 2$).

3.1. SCA

3.1.1. Formal description

Simultaneous component analysis [8] is an extension of principal component analysis to the multiset case. Different variants of SCA have been proposed that differ in the way the data blocks are scaled prior to the actual simultaneous component analysis; see [9] for an overview. A key element is that the different data blocks are represented by the same mathematical structure: In case of objectwise linked data $\mathbf{X}_k$ of size $I \times J_k$ the model for R simultaneous components becomes

$$\mathbf{X}_k = \mathbf{T}\mathbf{P}_k^{\mathrm{T}} + \mathbf{E}_k \quad \text{for all } k, \tag{1}$$

with $\mathbf{T}$ the $I \times R$ matrix of component scores that is shared between all blocks and $\mathbf{P}_k$ the $J_k \times R$ matrix of component loadings for block k; and, in case of variable-wise linked data $\mathbf{X}_k$ of size $I_k \times J$ it becomes

$$\mathbf{X}_k = \mathbf{T}_k\mathbf{P}^{\mathrm{T}} + \mathbf{E}_k \quad \text{for all } k, \tag{2}$$

with $\mathbf{T}_k$ the $I_k \times R$ matrix of component scores for block k and $\mathbf{P}$ the $J \times R$ matrix of component loadings shared between all blocks. Usually an orthogonality constraint is imposed on the component score matrices: either $\mathbf{T}^{\mathrm{T}}\mathbf{T} = \mathbf{I}$ (model (1)) or $\Sigma_k \mathbf{T}_k{}^{\mathrm{T}}\mathbf{T}_k = \mathbf{I}$ (model (2)); other types of constraints for the case of variable-wise linked data have been discussed by [18]. To simultaneously estimate all component score and loading matrices, a least squares criterion is used:

$$\min_{\mathbf{T},\mathbf{P}_k} \sum_k \left\| \mathbf{X}_k - \mathbf{T}\mathbf{P}_k{}^{\mathrm{T}} \right\|^2 \quad \text{such that } \mathbf{T}^{\mathrm{T}}\mathbf{T} = \mathbf{I}, \tag{3}$$

with $\|\mathbf{Z}\|^2$ indicating the sum of squared elements of the matrix $\mathbf{Z}$. In case of variable-wise linked data the objective is

$$\min_{\mathbf{T}_k,\mathbf{P}} \sum_k \left\| \mathbf{X}_k - \mathbf{T}_k\mathbf{P}^{\mathrm{T}} \right\|^2 \quad \text{such that } \sum\nolimits_k \mathbf{T}_k{}^{\mathrm{T}}\mathbf{T}_k = \mathbf{I}. \tag{4}$$

Objective functions (3) and (4) show that SCA heavily depends on the sum-of-squares of the data $\mathbf{X}_k$: This implies that variables with an offset that is high, in absolute value, or that have a large spread will dominate the solution and also that data blocks with a high sum-of-squares dominate the solution. For example, when data blocks differ considerably in size, the larger data block may dominate the solution. Such effects may be accounted for by mean-centering and scaling to equal sum-of-squares per block [19]; see [18] for a discussion on the matter of centering variables either per block or over the blocks for the case of variable-wise linked data and [9] on the matter of block-scaling. A solution to Eqs. (3) and (4) can be obtained by subjecting the concatenated data matrices $[\mathbf{X}_1 \ldots \mathbf{X}_k]$, respectively $[\mathbf{X}_1^{\mathrm{T}} \ldots \mathbf{X}_k{}^{\mathrm{T}}]^{\mathrm{T}}$, to a singular value decomposition and putting the component scores equal to the R left singular vectors associated to the R largest singular values and the loadings to the corresponding R right singular vectors multiplied by the R largest singular values. For this solution it also holds that the loadings are orthogonal: for model (3) this is $[\mathbf{P}_1^{\mathrm{T}}\ \mathbf{P}_2^{\mathrm{T}}]\,[\mathbf{P}_1^{\mathrm{T}}\ \mathbf{P}_2^{\mathrm{T}}]^{\mathrm{T}} = \mathbf{D}^2$ and for model (4) $\mathbf{P}^{\mathrm{T}}\mathbf{P} = \mathbf{D}^2$, with $\mathbf{D}^2$ a diagonal matrix. Note that there is no unique solution to Eqs. (3) and (4); for example, the orthogonal rotation of $\mathbf{T}$ by $\mathbf{B}$ with $\mathbf{B}^{\mathrm{T}}\mathbf{B} = \mathbf{B}\mathbf{B}^{\mathrm{T}} = \mathbf{I}$ can be compensated by counter-rotating the loadings $\mathbf{P}_k$ and preserves orthogonality of the component scores ($\mathbf{B}^{\mathrm{T}}\mathbf{T}^{\mathrm{T}}\mathbf{T}\mathbf{B} = \mathbf{I}$) but not of the loadings.

The fit of the different model components in Eq. (1) to a data block k can be assessed by the proportion of variation accounted for (VAF) by component r:

$$\left(1 - \left\| \mathbf{X}_k - \mathbf{t}_r\mathbf{p}_{kr}{}^{\mathrm{T}} \right\|^2\right) / \|\mathbf{X}_k\|^2, \tag{5}$$

and in case of model (2) this becomes

$$\left(1 - \left\| \mathbf{X}_k - \mathbf{t}_{kr}\mathbf{p}_r{}^{\mathrm{T}} \right\|^2\right) / \|\mathbf{X}_k\|^2. \tag{6}$$

However, caution is needed when interpreting the results from Eqs. (5) and (6) as the VAF in a data block k by a component r. The interpretation only holds when either the component scores or loadings are orthogonal at the level of the blocks. Else, the VAF by a component to a block also depends on the other components and, from a conceptual point of view, is not the VAF by component r. That orthogonality is needed in order to have pure contributions by a component to the VAF, independent from the other components, can be derived from the decomposition of the VAF jointly by the R components. For ease of demonstration we explicitly use $\|\mathbf{X}_k\|^2 = 1$. Then, the VAF for variable-wise linked data can be elaborated as follows:

$$\begin{aligned} 1 - \left\| \mathbf{X}_k - \mathbf{T}_{kr}\mathbf{P}^T \right\|^2 &= 1 - \left\| \mathbf{X}_k - \textstyle\sum_r \mathbf{t}_{kr}\mathbf{p}_r{}^{\mathrm{T}} \right\|^2 \\ &= 1 - \mathrm{tr}[\left(\mathbf{X}_k - \textstyle\sum_r \mathbf{t}_{kr}\mathbf{p}_r{}^{\mathrm{T}}\right)^{\mathrm{T}}\left(\mathbf{X}_k - S_r\mathbf{t}_{kr}\mathbf{p}_r{}^{\mathrm{T}}\right)] \\ &= 1 - \mathrm{tr}[\mathbf{X}_k{}^{\mathrm{T}}\mathbf{X}_k + \left(\textstyle\sum_r \mathbf{t}_{kr}\mathbf{p}_r{}^{\mathrm{T}}\right)^T\left(\textstyle\sum_r \mathbf{t}_{kr}\mathbf{p}_r{}^{\mathrm{T}}\right) - 2\mathbf{X}_k{}^T\left(\textstyle\sum_r \mathbf{t}_{kr}\mathbf{p}_r{}^{\mathrm{T}}\right)] \\ &= 2tr\mathbf{X}_k{}^T\left(\textstyle\sum_r \mathbf{t}_{kr}\mathbf{p}_r{}^{\mathrm{T}}\right) - \mathrm{tr}\left(\textstyle\sum_r \mathbf{t}_{kr}\mathbf{p}_r{}^{\mathrm{T}}\right)^T\left(\textstyle\sum_r \mathbf{t}_{kr}\mathbf{p}_r{}^{\mathrm{T}}\right). \end{aligned} \tag{7}$$

Please cite this article as: K. Van Deun, et al., Identifying common and distinctive processes underlying multiset data, Chemometrics and Intelligent Laboratory Systems (2013), http://dx.doi.org/10.1016/j.chemolab.2013.07.005

Compare this with the sum of the VAF per component,

$$\begin{aligned}\sum_r\left(1-\left\|\mathbf{X}_k-\mathbf{t}_{kr}\mathbf{p}_r^{\mathrm{T}}\right\|^2\right) &= R-\sum_r\left\|\mathbf{X}_k-\mathbf{t}_{kr}\mathbf{p}_r^{\mathrm{T}}\right\|^2\\ &= R-\sum_r\left\{\mathrm{tr}\left[\left(\mathbf{X}_k-\mathbf{t}_{kr}\mathbf{p}_r^{\mathrm{T}}\right)^{\mathrm{T}}\left(\mathbf{X}_k-\mathbf{t}_{kr}\mathbf{p}_r^{\mathrm{T}}\right)\right]\right\}\\ &= R-\mathrm{tr}\left[\sum_r\mathbf{X}_k^{\mathrm{T}}\mathbf{X}_k+\sum_r\mathbf{p}_r\mathbf{t}_{kr}^{\mathrm{T}}\mathbf{t}_{kr}\mathbf{p}_r^{\mathrm{T}}-2\mathbf{X}_k^{\mathrm{T}}\left(\sum_r\mathbf{t}_{kr}\mathbf{p}_r^{\mathrm{T}}\right)\right]\\ &= 2tr\mathbf{X}_k^{\mathrm{T}}\left(\sum_r\mathbf{t}_{kr}\mathbf{p}_r^{\mathrm{T}}\right)-\mathrm{tr}\left(\sum_r\mathbf{p}_r\mathbf{t}_{kr}^{\mathrm{T}}\mathbf{t}_{kr}\mathbf{p}_r^{\mathrm{T}}\right).\end{aligned} \tag{8}$$

This shows that the VAF jointed by R components can be decomposed in contributions per component if in Eq. (7) the term $\mathrm{tr}(\Sigma_r\mathbf{t}_{kr}\mathbf{p}_r^T)^T(\Sigma_r\mathbf{t}_{kr}\mathbf{p}_r^{\mathrm{T}}) = \mathrm{tr}(\Sigma_r\mathbf{p}_r\mathbf{t}_{kr}^T\mathbf{t}_{kr}\mathbf{p}_r^{\mathrm{T}})$; this is the case when either $\mathbf{P}$ or $\mathbf{T}_k$ is orthogonal.

3.1.2. Application to the illustrative data

3.1.2.1. Objectwise linked data. We apply SCA to the data obtained for the group of 24 TIV vaccinees three and seven days after vaccination. Note that these data were pre-processed: each variable was centered and scaled to sum-of-squares one with the outcome that the data blocks have equal sum-of-squares given that they have the same size (namely $J_1 = J_2 = 54{,}715$). In this way, each gene and each data block receive the same weight in the analysis. The bars in Fig. 2 display the proportion of VAF by the simultaneous components in the two data blocks (upper panel: Day 3; middle panel: Day 7) as well as in the concatenated data (lower panel). For Day 3, the first four components seem to stand somewhat out while for Day 7 this is the case for the first (five) component(s). Given our interest in both common and specific components, we also retain components that may be specific for a particular block; hence, we retain the five component solution ($R = 5$). The VAF by each component in each data block is shown also in Table 1; the simultaneous components account for similar amounts of variance in each of the data blocks hence they may be considered common.

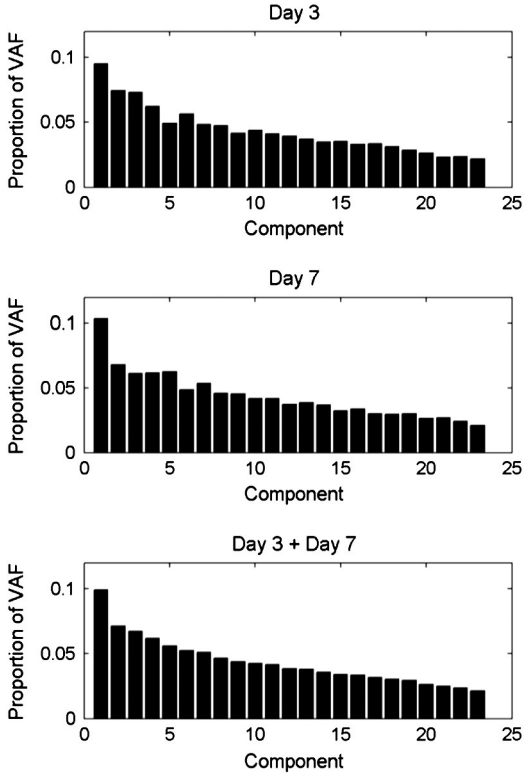

Fig. 2. Proportion of variance accounted for by the simultaneous components for each data block individually (top panel: Day 3; middle panel: Day 7) and the concatenated data (bottom panel).

Table 1 (under the header 'SCA') shows the VAF per component in each data block and in the concatenated data for the five selected components. It also shows the total VAF by the five components simultaneously and, because the matrix of component scores is orthogonal for each data block, the total VAF equals the sum of the componentwise VAF. Note that the simultaneous components are sequentially optimal with respect to the VAF in the concatenated data (hence the decreasing VAF in the column 'Conc.' of Table 1). In Table 2, the correlation of the five components with the antibody titers is presented, the sum of the squared correlations, and the coefficient of determination R^2 as obtained from a regression of the antibody titers on the five components jointly. Note that the simultaneous component scores in Eq. (1) are orthogonal with the result that R^2 equals the sum of the squared correlations.

The same underlying matrix of component scores is used to model both Day 3 and Day 7 data. This matrix reflects interindividual differences in susceptibility for the vaccine. Given that the components contribute for very similar amounts of variation in both data blocks, they can be considered to be common components underlying both Day 3 and Day 7 data. This reflects that interindividual differences in response to the vaccine barely vary over time. From an immunological point of view this makes sense, little differences are to be expected in the inter-individual variation of the immune response three or seven days after vaccination: individuals that are responsive to the vaccine at Day 3 are still so at Day 7 and the same is true for non-responders. The second components predicts best the antibody titers. To find the biological functions associated to the components, we rely on the GSEA [16] functional annotation tool. This tool requires a ranking of the genes with respect to their importance for the component; such information is contained in the loadings. A ranking of the genes based on the Day 3 associated loadings on the second component was subjected to GSEA, as well as one for the Day 7 associated loadings. The revealed sets are those based on [1] and refer to genes differentially expressed in persons vaccinated with TIV against control. The content of the terms as well as the fact that the same terms are found for the two sets of loadings, fits the expectation of shared processes.

3.1.2.2. Variable-wise linked data. The expression data pertaining to the same set of 54,715 probe sets obtained from subjects vaccinated either with TIV or with LAIV, three days after vaccination, were mean-centered and scaled to sum-of-squares one per probeset for each block. Compared to centering and scaling the variables over the blocks, this removes differences in means between the blocks and also differences in variability that may exist between blocks. It also results in an equal sum of squares per data block such that the blocks receive the same weight in the simultaneous component analysis. The fit of the different components to the TIV and LAIV data blocks, and the concatenated data is depicted in Fig. 3. As we used the unrotated SCA solution the loadings are orthogonal; this allows to calculate pure contributions of the components to the VAF at the level of the blocks. The first and third components stand out in the TIV block; in the LAIV block the first six components stand out. We select the six component solution as we are also interested in components that are specific for a single block. Although each of the six simultaneous components underlies the concatenated data, there may be differences between the blocks in the VAF by a particular component. For example, the third component accounts for 7% of the variation in the TIV block and only for 3% in the LAIV block (see Table 3, under the header 'SCA'). Table 3 also shows that the sum of VAF is equal to the total VAF; this is due to the orthogonality of the loadings per block. However, within blocks the component scores are not orthogonal. Therefore, the coefficient of determination R^2 is not equal to the sum of the squared correlations of the block-specific component scores with the antibody titers (compare the two last lines in Table 4 under the header 'SCA').

In this case of variable-wise linked data, the same matrix of loadings applies to both data blocks while the component scores pertain to

Please cite this article as: K. Van Deun, et al., Identifying common and distinctive processes underlying multiset data, Chemometrics and Intelligent Laboratory Systems (2013), http://dx.doi.org/10.1016/j.chemolab.2013.07.005

Table 1
Proportion of variation accounted for by each of five (pairs of) components, their sum, and jointly (the lines C1–C5, Sum, and Total respectively) in each data block (Day 3 or Day 7) by the six methods (SCA, DISCO-SCA, adapted GSVD, ECO-POWER, O2PLS, and CCA). For the first four methods, also the proportion of variation accounted for in the concatenated data is reported. A color code is used to indicate the status of the component: Yellow for Day 3 distinctive components, pink for Day 7 distinctive components, and green for common components; because the status of the SCA components is unspecified no color was used in the corresponding cells.

	SCA			DISCO-SCA			Adapted GSVD			ECO-POWER			CCA		O2PLS	
	Day 3	Day 7	Conc.	Day 3	Day 7	Conc.	Day 3	Day 7	Conc.	Day 3	Day 7	Conc.	Day 3	Day 7	Day 3	Day 7
C1	0.09	0.10	0.10	0.09	0.10	0.10	0.09	0.05	0.07	0.10	0.10	0.10	0.10	0.11	0.10	0.11
C2	0.07	0.07	0.07	0.07	0.07	0.07	0.09	0.06	0.08	0.07	0.06	0.07	0.07	0.07	0.07	0.07
C3	0.07	0.06	0.07	0.07	0.06	0.07	0.06	0.07	0.07	0.07	0.07	0.07	0.07	0.06	0.07	0.06
C4	0.06	0.06	0.06	0.06	0.06	0.06	0.07	0.09	0.08	0.05	0.06	0.06	0.06	0.06	0.06	0.06
C5	0.05	0.06	0.06	0.05	0.06	0.06	0.05	0.08	0.06	0.07	0.06	0.07	0.05	0.06	0.06	0.04
Sum	0.35	0.36	0.36	0.35	0.36	0.36	0.35	0.36	0.36	0.35	0.36	0.36	0.36	0.37	0.36	0.34
Total	0.35	0.36	0.36	0.35	0.36	0.36	0.35	0.36	0.36	0.35	0.36	0.36	0.36	0.37	0.36	0.34

the individuals of a specific block. This means that at the gene level, the underlying process is the same for both blocks. However, the interindividual variability in these processes may differ between the blocks. From an immunological point of view this makes sense as different vaccines may have a different impact on immunity-related processes and thus induce more or less or even no variability with respect to these processes. At this point, we would like to stress that differences in means between the vaccines were removed by centering the data per data block. The components are better at predicting the antibody titers for the vaccinees treated with TIV and there is a strong correlation ($r = 0.64$) between the scores on the third component and the titers. Subjecting a ranking of the genes based on the loadings of this component to GSEA, pointed towards genes downregulated in CD4 and CD8 T cells compared to monocytes and myeloid and upregulated after vaccination with TIV or yellow fever vaccine YF-17D compared to control samples. The latter is a live attenuated vaccine (LA) and "one of the most successful vaccines ever developed" [1].

3.2. DISCO-SCA

3.2.1. Formal description

DISCO-SCA [13,14] is a method that exploits the rotational freedom of the simultaneous components resulting from the optimization of Eqs. (3) and (4) and rotates the components to a DIStinctive and COmmon structure. The rotation criterion used is a partially specified target criterion where the target is one that specifies distinctive components as components having zero scores (loadings) in the positions that correspond to the data block the component do not underlie. All remaining entries of the target are arbitrary. In case of objectwise linked data, the rotation matrix $\mathbf{B}$ is found by minimizing

$$\min(\mathbf{B}) \left\| \mathbf{W}^{\circ} \left(\mathbf{P}_{target} - \left[\mathbf{P}_1{}^{\mathrm{T}} \mathbf{P}_2{}^{\mathrm{T}} \right]^{\mathrm{T}} \mathbf{B} \right) \right\|^2 \text{such that } \mathbf{B}^{\mathrm{T}}\mathbf{B} = \mathbf{I} = \mathbf{B}\mathbf{B}^{\mathrm{T}}, \quad (9)$$

with $\mathbf{W}$ a binary matrix having ones in the positions corresponding to the specified entries in the target and zeroes elsewhere; $^{\circ}$ denotes the elementwise product. When the data are linked variable-wise, the target is specified for the component scores yielding

$$\min(\mathbf{B}) \left\| \mathbf{W}^{\mathrm{t}} \left(\mathbf{T}_{target} - \left[\mathbf{T}_1{}^{\mathrm{T}} \mathbf{T}_2{}^{\mathrm{T}} \right]^{\mathrm{T}} \mathbf{B} \right) \right\|^2 \text{such that } \mathbf{B}^{\mathrm{T}}\mathbf{B} = \mathbf{I} = \mathbf{B}\mathbf{B}^{\mathrm{T}}. \quad (10)$$

Note that in Eqs. (9) and (10) the number of components and their status (distinctive for $\mathbf{X}_1$, distinctive for $\mathbf{X}_2$, or common) has to be prespecified, which can be considered a model selection issue.

3.2.2. Application to the illustrative data

3.2.2.1. Objectwise linked data. We applied DISCO-SCA to the Day 3 and Day 7 data obtained for the same set of subjects. Regarding model selection, as a first step in the analysis, the loading matrix obtained by SCA was rotated to every possible target for a solution with five components (five components because this is the number selected in the SCA

Table 2
Correlation of the components with the antibody titers, their sum of squares, and total R^2 for the multiple regression of the titers on the five components jointly. A color code is used to indicate the status of the component: Yellow for Day 3 distinctive components, pink for Day 7 distinctive components, and green for common components; because the status of the SCA components is unspecified no color was used in the corresponding cells.

	SCA	DISCO-SCA	Adapted GSVD	ECO-POWER	CCA		O2PLS	
					Day 3	Day 7	Day 3	Day 7
C1	−0.40	−0.40	0.40	−0.38	0.38	0.42	−0.35	−0.39
C2	−0.45	−0.45	0.35	−0.24	0.43	−0.31	0.38	0.27
C3	0.29	0.29	0.12	0.45	0.33	−0.32	−0.29	−0.26
C4	0.10	0.10	0.34	0.22	0.09	0.13	−0.21	−0.22
C5	0.12	0.12	−0.31	−0.15	0.11	0.15	−0.21	−0.20
Sum R^2	0.47	0.47	0.51	0.47	0.45	0.41	0.44	0.38
R^2	0.47	0.47	0.47	0.47	0.45	0.41	0.44	0.38

Please cite this article as: K. Van Deun, et al., Identifying common and distinctive processes underlying multiset data, Chemometrics and Intelligent Laboratory Systems (2013), http://dx.doi.org/10.1016/j.chemolab.2013.07.005

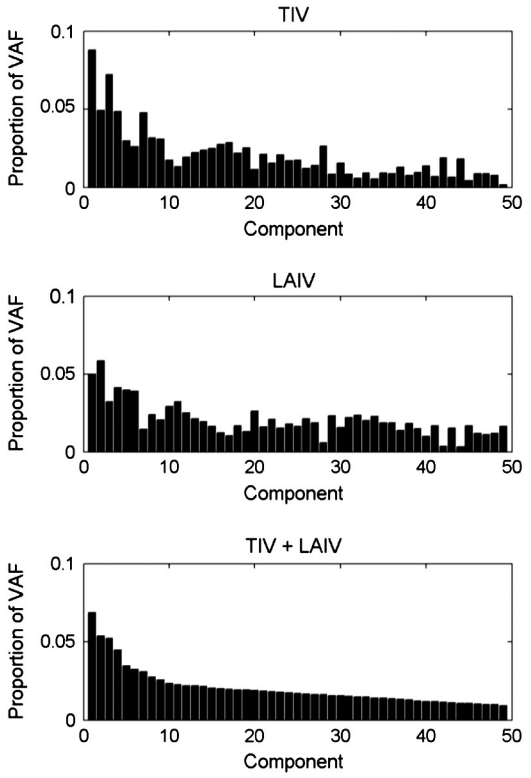

Fig. 3. Proportion of variance accounted for by the simultaneous components for each data block individually (top panel: TIV; middle panel: LAIV) and the concatenated data (bottom panel).

analysis). For each target a deviation score was obtained. This score is the maximum of the componentwise deviations calculated as follows: 1) for a distinctive component the deviation score equals the sum of squares of the loadings targeted to be zero, divided by the sum of squares of the data block, and, 2) for a common component as the difference between the block specific sum of squares divided by the sum of squares of the data block [14]. In Fig. 4 the deviation score of each target is plotted against the number of distinctive components that underlies the target. The solution with the lowest deviation was retained; this was the solution with all common components which coincides with the SCA solution. We refer to Section 3.1.2.1 for further discussion of the results.

3.2.2.2. Variable-wise linked data. The plot of deviation scores obtained for DISCO-SCA applied to the expression data obtained from the TIV and LAIV group three days after vaccination is shown in Fig. 5. Here, we rotated to all possible targets with six components. Furthermore, the componentwise deviation scores were now calculated on the component scores rather than on the loadings. The lowest deviation was obtained for the solution with three common components and one component that is distinctive for LAIV and two that are distinctive for TIV. The proportion of VAF by the components in each data block is shown in Table 3, in the columns headed by 'DISCO-SCA'. Compared to the VAF by the SCA components, the status of the components is now much clearer. Within blocks, the components are not orthogonal; therefore the sum of VAF is (slightly) different from the VAF contributed by the six components.

The interpretation of a 'distinctive' component is that it is a component that is clearly absent in a particular block. For example, the distinctive component for TIV does not account for a substantial amount of variance in the LAIV data and the processes it may represent are consequently absent in the LAIV data. The fact that the same loadings are used for both data blocks is crucial in this respect: the interpretation of the component is one and the same for both data blocks but the interindividual variation (and resulting VAF) is only considerable in the TIV data block. In part, this is confirmed by the results in Table 4: The second distinctive component for TIV strongly correlates with the TIV titers only while the third common component correlates with the antibody titers for both group of vaccinees (TIV and LAIV).

Component 1 is specific for the LAIV data block and it is the component with the highest VAF for this data block; note, however, that this component has a low correlation with the antibody titers for the LAIV treated subjects. An annotation of the component reveals sets of genes that are upregulated in LAIV compared to control subjects and in CD4 and CD8 T cells compared to monocytes, neutrophils and myeloid cells. Also, sets of genes downregulated in patients that acquired viral or bacterial infections were found.

The third component is distinctive for the TIV data block and accounts for almost no variation in the LAIV data. It has a very strong correlation with the titers ($r = 0.62$). Subjecting the loadings to GSEA yields sets referring to genes down-regulated in CD4 and CD8 T cells and in B cells compared to monocytes, dendritic, and myeloid cells and upregulated in memory CD4 T cells, in patients having acquired a viral infection, and in subjects vaccinated with YF-17D. Also the sets based on the differential expression analysis of the TIV against control [1] are recovered by sets of genes that are upregulated in TIV.

The sixth component is shared and accounts for variation in both the TIV block and the LAIV block. Although this is a common component represented by the same set of loadings (there is only one set of loadings

Table 3
Proportion of variation accounted for by each of six components, their sum, and jointly (the lines C1–C6, Sum, and Total respectively) in each data block (TIV or LAIV) by four methods (SCA, DISCO-SCA, adapted GSVD, and O2PLS). For the first three methods, also the proportion of variation accounted for in the concatenated data is reported. A color code is used to indicate the status of the component: Yellow for TIV distinctive components, pink for LAIV distinctive components, and green for common components; because the status of the SCA components is unspecified no color was used in the corresponding cells.

	SCA			DISCO-SCA			Adapted GSVD			O2PLS	
	TIV	LAIV	Conc.	TIV	LAIV	Conc.	TIV	LAIV	Conc.	TIV	LAIV
C1	0.09	0.05	0.07	0.02	0.06	0.04	0.08	0.02	0.05	0.10	0.08
C2	0.05	0.06	0.05	0.09	0.03	0.06	0.09	0.03	0.06	0.07	0.07
C3	0.07	0.03	0.05	0.08	0.02	0.05	0.05	0.04	0.04	0.08	0.06
C4	0.05	0.04	0.04	0.05	0.05	0.05	0.05	0.05	0.05	0.03	0.03
C5	0.03	0.04	0.03	0.04	0.05	0.05	0.03	0.05	0.04	0.04	0.03
C6	0.03	0.04	0.03	0.03	0.04	0.03	0.02	0.06	0.04	0.04	0.03
Sum	0.31	0.26	0.29	0.31	0.26	0.29	0.31	0.26	0.29	0.36	0.30
Total	0.31	0.26	0.29	0.31*	0.26*	0.29	0.31	0.26	0.29	0.36	0.30

*Sum of componentwise VAF differs from total VAF (difference of 5.9e−06 for TIV and of −5.9e−06 for LAIV).

Please cite this article as: K. Van Deun, et al., Identifying common and distinctive processes underlying multiset data, Chemometrics and Intelligent Laboratory Systems (2013), http://dx.doi.org/10.1016/j.chemolab.2013.07.005

Table 4
Correlation of the components with the antibody titers, their sum of squares, and total effect size R^2 for the regression of the titers on the six components obtained with four different methods (SCA, DISO-SCA, adapted GSVD, and O2PLS). The titers for the subjects vaccinated with the TIV vaccine and the LAIV vaccine were treated separately. A color code is used to indicate the status of the component: Yellow for TIV distinctive components, pink for LAIV distinctive components, and green for common components; because the status of the SCA components is unspecified no color was used in the corresponding cells.

	SCA		DISCO-SCA		Adapted GSVD		O2PLS	
	TIV	LAIV	TIV	LAIV	TIV	LAIV	TIV	LAIV
C1	0.22	0.05	−0.20	0.06	0.59	−0.19	−0.20	0.06
C2	−0.01	0.03	−0.01	0.03	0.18	−0.07	−0.29	-0.03
C3	0.64	−0.07	0.62	−0.20	0.32	−0.29	−0.53	0.03
C4	0.16	−0.23	−0.33	−0.10	−0.24	−0.17	0.13	-0.26
C5	0.41	−0.37	0.11	0.07	0.09	−0.37	0.23	-0.02
C6	−0.08	−0.28	−0.25	0.49	−0.20	0.06	−0.45	0.21
Sum R^2	0.65	0.28	0.61	0.30	0.60	0.29	0.68	0.12
R^2	0.60	0.29	0.60	0.29	0.60	0.29	0.59	0.12

to represent both data blocks), the component correlates negatively with the antibody titers for subjects treated with TIV but positively for those treated with LAIV. An enrichment analysis based on a ranking of the genes on this component results in gene sets that are downregulated in subjects vaccinated with yellow fever LA vaccine or with LAIV flu vaccine. Also, sets of genes were found that are downregulated in memory CD8 T cells compared to effector CD8 T cells.

3.3. Adapted GSVD

3.3.1. Formal description

In the field of computational biology, the Generalized Singular Value Decomposition (GSVD) is a popular method for finding common and distinctive processes [12]. [14,20] showed the need to adapt the GSVD matrix decomposition in order to become a suitable dimension reduction method. For objectwise linked data, the data are modeled as

$$\begin{aligned}\mathbf{X}_k &= \mathbf{T}\mathbf{D}_k\mathbf{V}_k^{\mathrm{T}} + \mathbf{E}_k \qquad \text{for} \quad k = 1,2,\\ &= \mathbf{T}\mathbf{P}_k^{\mathrm{T}} + \mathbf{E}_k\end{aligned} \tag{11}$$

with $\mathbf{V}_k^{\mathrm{T}}\mathbf{V}_k = \mathbf{I}$ and $\mathbf{D}_k$ diagonal and such that $\mathbf{D}_1^2 + \mathbf{D}_2^2 = \mathbf{I}$; for variable-wise data the adapted GSVD models the data as

$$\begin{aligned}\mathbf{X}_k &= \mathbf{U}_k\mathbf{D}_k\mathbf{P}^{\boldsymbol{T}} + \mathbf{E}_k \qquad \text{for} \quad k = 1,2,\\ &= \mathbf{T}_k\mathbf{P}^{\mathrm{T}}\end{aligned} \tag{12}$$

with $\mathbf{U}_k^{\mathrm{T}}\mathbf{U}_k = \mathbf{I}$. The status of the components can be derived from the diagonal matrices: if $d_{1R}^2 \approx 1$, the component is distinctive for $\mathbf{X}_1$, if $d_{2R}^2 \approx 1$, the component is distinctive for $\mathbf{X}_2$, and if $d_{1R}^2 \approx d_{1R}^2 \approx 0.50$, the component is common. To our knowledge, there is no explicit objective function underlying the GSVD. The algorithm of Paige and Saunders [21] in its adapted form [14] consists of two steps. The first step is a simultaneous component analysis, namely a singular value decomposition of the concatenated data. The second step uses the resulting right (left) singular vectors for objectwise (variable-wise) linked data respectively and performs a SVD for each of the two block-specific parts in these matrices. The singular values of the SVDs in the second step correspond to $\mathbf{D}_1$ and $\mathbf{D}_2$ in Eqs. (11) and (12), respectively. As shown by [14,21] the latter step is a non-singular transformation of the SCA solution obtained in the first step and transforms the components to components that account for maximal variation in $\mathbf{X}_1(\mathbf{X}_2)$ relative to $\mathbf{X}_2(\mathbf{X}_1)$. Nevertheless, the (adapted) GSVD can be understood as a two-step procedure with the first step being a simultaneous component analysis and the second step one that uses the non-uniqueness of SCA to find distinctive components. Therefore, the adapted GSVD is a least-squares method with maximal VAF in the concatenated data and, in case of variable-wise linked data, it is a rotation of SCA (see [14]) and it gives a closed form solution for the SCA-IND model [18]. The difference between the adapted GSVD and the regular GSVD resides in the rank reduction: In the algorithm of the adapted GSVD only the R highest

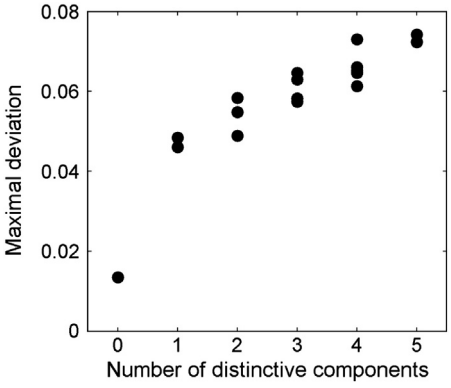

Fig. 4. Day 3–Day 7 multiset data: For each possible target, the maximal deviation is plotted in function of the number of distinctive components.

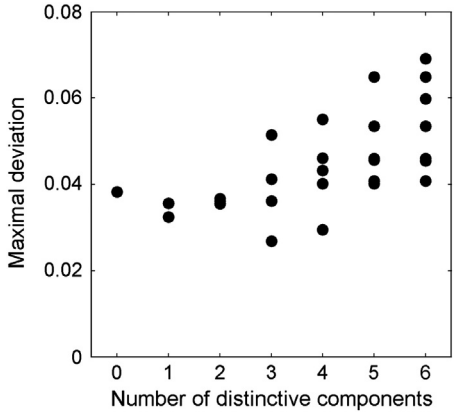

Fig. 5. TIV-LAIV multiset data: For each possible target, the maximal deviation is plotted in function of the number of distinctive components.

Please cite this article as: K. Van Deun, et al., Identifying common and distinctive processes underlying multiset data, Chemometrics and Intelligent Laboratory Systems (2013), http://dx.doi.org/10.1016/j.chemolab.2013.07.005

singular values and corresponding vectors are retained for further analysis in step 2 while in the original algorithm of Paige and Saunders all non-zero singular values and associated vectors are retained.

3.3.2. Application to the illustrative data

3.3.2.1. Objectwise linked data. The adapted GSVD was applied to the data obtained three and seven days after vaccination with TIV. We selected the same number of components as before, namely five components. Table 1 contains the proportion of VAF by the components in each data block under the header 'Adapted GSVD': The first component accounts for a considerable amount of variance in the Day 3 block and for almost no variance in the Day 7 block, and, hence, can be considered a Day 3 distinctive component; the second, third, and fourth components account for a reasonable amount in both data blocks, and therefore are common components; the last component accounts for a lot of variation in the Day 7 data and for almost no variation in the Day 3 data, this is a Day 7 distinctive component. This illustrates that compared to (DISCO-)SCA the GSVD is oriented to find components that account for much more variance in one data block compared to the other. From Table 1, it can also be observed that the VAF accounted for by the five components in a particular block can be obtained from the contributions by the individual components; this is the case because the GSVD loadings are orthogonal within blocks (see Eq. (11)). The GSVD component scores, on the other hand, are not orthogonal: hence the sum of squared correlations of the components with the titers is not equal to the R^2 obtained from the regression of the titers on all components simultaneously (see Table 2).

The model structure in Eq. (11) that relates to the transcripts, are the block-specific $\mathbf{P}_k = \mathbf{V}_k\mathbf{D}_k$ matrices. Component 1 has the highest correlation with the antibody titers ($r = 0.40$). The annotation of the associated loadings yields sets of genes upregulated in monocytes compared to T- (CD4 and CD8) and B-cells and in subjects vaccinated with TIV or yellow fever vaccine.

3.3.2.2. Variable-wise linked data. Application of the adapted GSVD to the data obtained three days after vaccination from the TIV and LAIV groups with $R = 6$, yields a solution with two TIV distinctive, one LAIV distinctive, and three common components (see Table 3). Especially the first TIV distinctive component correlates highly with the antibody titers for the subjects vaccinated with TIV. In this case, the structure related to the genes is the matrix $\mathbf{P}$ in Eq. (12). The first component is distinctive for the TIV data block and accounts for almost no variation in the LAIV data. Subjecting the loadings to GSEA yields basically the same sets as found for the DISCO-SCA TIV distinctive component except for the sets related to memory CD4 T cells that are not revealed by the adapted GSVD. The third component is shared between the two data blocks but correlates positively with the antibody titers for the vaccinees treated with TIV and negatively for those treated with LAIV. The same gene sets are found as for the common (sixth) DISCO-SCA component. Component 6 is specific for the LAIV data block and it is the component with the highest VAF for this data block; note, however, that this component has a low correlation with the antibody titers for the LAIV treated subjects. The annotation of the component resulted in the same sets of genes found for the DISCO-SCA LAIV distinctive component.

3.4. Eco-power

3.4.1. Formal description

Inspired by power regression [22], Schouteden et al. [10] introduced a simultaneous component method for objectwise linked data relying on model (1) and with the following objective function:

$$\max_{\mathbf{W}} \sum_r R_{r1}^2 R_{r2}^2 \quad \text{such that} \quad \mathbf{T}^{\mathrm{T}}\mathbf{T} = \mathbf{I}, \tag{13}$$

where $\mathbf{T} = [\mathbf{X}_1\ \mathbf{X}_2]\mathbf{W}$ with $\mathbf{W}$ being of size $(J_1 + J_2) \times R$ (which implies that the component scores lie in the space spanned by the concatenated data matrix $[\mathbf{X}_1\ \mathbf{X}_2]$), and with R_{rk}^2 the proportion of variance accounted for by component r in data block k,

$$R_{rk}{}^2 = \left(\|\mathbf{X}_k\|^2\right)^{-1} \mathbf{w}_r{}^{\mathrm{T}}[\mathbf{X}_1\mathbf{X}_2]^{\mathrm{T}}\mathbf{X}_k\mathbf{X}_k{}^{\mathrm{T}}[\mathbf{X}_1\mathbf{X}_2]\mathbf{w}_r \quad \text{for} \quad k = 1, 2. \tag{14}$$

The motivation to take the product of the proportion of VAF, is to obtain components that are *common* in the sense that they account for a similar amount of variance in both data blocks and this for each component (hence the product is taken per component). In Eq. (14) it is the proportion of VAF that is taken as shown by the scale factor $(\|\mathbf{X}_k\|^2)^{-1}$ which is the inverse of the sum of squares of the data block. This means that ECO-POWER gives equal weight to each data block, independent of their relative sizes and scale. To find a solution to Eq. (14), [10] rely on an iterative majorization procedure. Given a solution for the component scores $\mathbf{T}$, loading matrices can be obtained as follows: $\mathbf{P}_1 = \mathbf{X}_1^{\mathrm{T}}\mathbf{T}^{\mathrm{T}}$ and $\mathbf{P}_2 = \mathbf{X}_2^{\mathrm{T}}\mathbf{T}^{\mathrm{T}}$.

3.4.2. Application to the illustrative data

3.4.2.1. Objectwise linked data. We used ECO-POWER to extract five components from the data. In Table 1 the proportion of VAF by these components in each data block is shown; for each component it holds that the VAF is almost exactly the same for the two data blocks as it is the objective of ECO-POWER to find such common components. Note that the components contribute in a pure way to the VAF for a data block (as the component scores are orthogonal). Together, these five components account for 47% of the variation in the antibody titers; components one and three account for the largest portion of this variation (see Table 2). The annotation of these components is based on a ranking of the genes by the matrices $\mathbf{P}_k$ and, for the third component yields 1) for the Day 3 data terms related to genes that are upregulated in monocytes and myeloids compared to CD4 and CD8 T cells and B-cells, and 2) for the Day 7 data, in addition terms were found associated to genes that are upregulated for subjects vaccinated against the flu and downregulated in memory B cells.

3.5. CCA

3.5.1. Formal description

Canonical correlation analysis is a method for the analysis of two-set data that are linked objectwise (with generalized canonical correlation analysis being the extension to multiset data). The model used is

$$\begin{aligned}\mathbf{X}_k &= \mathbf{X}_k\mathbf{W}_k\mathbf{P}_k{}^{\mathrm{T}} + \mathbf{E}_k \\ &= \mathbf{T}_k\mathbf{P}_k{}^{\mathrm{T}} + \mathbf{E}_k \quad \text{for} \quad k = 1, 2,\end{aligned} \tag{15}$$

with $\mathbf{W}_k$ of size $J_k \times R$ the matrix of canonical weight coefficients and $\mathbf{T}_k$ of size $I \times J_k$ the matrix of canonical variates which are further subject to $I^{-1}\mathbf{T}_k{}^{\mathrm{T}}\mathbf{T}_k = \mathbf{I}$. Interest is in maximizing the sum of the correlations between the R pairs of canonical variates as expressed by the following objective function:

$$\max(\mathbf{W}_1, \mathbf{W}_2)\ \mathrm{tr}\mathbf{W}_1{}^{T}\mathbf{X}_1{}^{T}\mathbf{X}_2\mathbf{W}_2 \text{ such that } I^{-1}\mathbf{T}_1{}^{\mathrm{T}}\mathbf{T}_1 = \mathbf{I} = I^{-1}\mathbf{T}_2{}^{\mathrm{T}}\mathbf{T}_2. \tag{16}$$

The canonical variates are block-specific and are defined as those linear combinations of the variables in a block that correlate maximally between blocks; the CCA solution is not influenced by the relative size and scale of the data blocks. The variance accounted for by the canonical variates in the data blocks is not taken into account in the objective function (16), and, hence, these variates may be poor summarizers of the data [23]. Note that different quantification matrices of the linking

Please cite this article as: K. Van Deun, et al., Identifying common and distinctive processes underlying multiset data, Chemometrics and Intelligent Laboratory Systems (2013), http://dx.doi.org/10.1016/j.chemolab.2013.07.005

mode, $\mathbf{T}_k$ result from the analysis, one for each data block, with $\mathbf{T}_k$ lying in the space spanned by $\mathbf{X}_k$. When $I < J_k$ for some k (Eq. (16)) yields an underdetermined system; to account for this problem, a regularized version of canonical correlation analysis can be used [24,25]. Let $\mathbf{R} = \Sigma_k \mathbf{X}_k(\mathbf{X}_k{}^T\mathbf{X}_k)^+\mathbf{X}_k{}^T$ with $\mathbf{Z}^+$ indicating the Moore–Penrose inverse of $\mathbf{Z}$; then a solution to Eq. (16) can be found using the eigenvectors of $\mathbf{R}$ (see [24]). Regularized CCA uses $\mathbf{R} = \Sigma_k \mathbf{X}_k[(1-\alpha)\mathbf{X}_k{}^T\mathbf{X}_k + \alpha\mathbf{I})]^+\mathbf{X}_k{}^T$ and $\alpha = 0$ results in ordinary CCA while $\alpha = 1$ yields SCA. When $a > 0$, the relative size and/or scale of the data blocks has an influence on the obtained solution as is the case for SCA. To account for such differences between blocks they may be scaled to equal sum of squares such that each block receives the same weight in the regularized canonical correlation analysis.

3.5.2. Application to the illustrative data

3.5.2.1. Objectwise linked data. The analyses were done with the RGCCA R package [26], and more specifically with its kernel extension[1] [27]. Note that the calculations for CCA are usually based on the matrices of cross-products between the variables; for the application here this is a matrix of size 54,715 by 54,715 which results in memory and computational problems, for example to calculate its inverse. The advice for such a high-dimensional problem is to set the regularization parameter equal to its maximal value (one).

Table 1 contains the proportions of VAF by the canonical variates (in the columns headed by 'CCA'). Most components account for a similar amount of variance in the two data blocks. As the component scores of a particular block are orthogonal, the components contribute in a pure way to the VAF in a block and hence the two last lines in Tables 1 and 3 are equal. Because of the kernel approach to regularized CCA, the results obtained are not exactly the same as those described by [24] and a solution different from the SCA solution is obtained. In particular, the block specific component scores $\mathbf{T}_k$ are not restricted to be the same for both data blocks and the total VAF for the Day 7 data is slightly more than with SCA. Note that regularized CCA may account for more variance in a particular block. The components of the two blocks, t_{1r} and t_{2r}, correlate highly for all r as this is the objective of CCA (the correlations ranged, in absolute value, from 0.95 to 0.98). The first two components have the highest correlation with the antibody titers. Their canonical weight matrices $\mathbf{W}_k$ in Eq. (15) were used to rank the genes in GSEA [16]. For the first component these were found to be significantly enriched for sets that are downregulated in T cells and in subjects vaccinated with LAIV while upregulated in subjects vaccinated with TIV. The second component is associated to genes upregulated in subjects vaccinated with TIV or LAIV flu vaccine or with yellow fever vaccine. These sets were found for both weight matrices (associated to the Day 3 and the Day 7 data).

3.5.2.2. Variable-wise linked data. CCA is specifically developed for the case of objectwise linked data. From a technical point of view it is possible to use the same methodology to analyze variable-wise linked data by transposing the data matrices. However, the canonical variates then obtain a strange status as these are the result of a linear combination of the samples.

3.6. O2PLS

3.6.1. Formal description

O2PLS [28] has been proposed in the chemometric literature as a method for the analysis of two objectwise linked data blocks with the aim of finding sources of shared variation between the data blocks and sources of variation that are specific for a data block. The data are modeled as follows,

$$\mathbf{X}_k = \mathbf{T}_{k,c}\mathbf{P}_{k,c}{}^T + \mathbf{T}_{k,s}\mathbf{P}_{k,s}{}^T + \mathbf{E}_k \quad \text{for all } k, \tag{17}$$

with $\mathbf{T}_{k,c}$ and $\mathbf{P}_{k,c}$ the shared component scores and loadings respectively, and $\mathbf{T}_{k,s}$ and $\mathbf{P}_{k,s}$ the distinctive component scores and loadings. The three parts of model (17) are orthogonal to each other: $\mathbf{E}_k^T(\mathbf{T}_{k,c}\mathbf{P}_{k,c}^T) = \mathbf{0}$, $\mathbf{E}_k^T(\mathbf{T}_{k,s}\mathbf{P}_{k,s}^T) = \mathbf{0}$, and $(\mathbf{T}_{k,s}\mathbf{P}_{k,s}^T)^T(\mathbf{T}_{k,c}\mathbf{P}_{k,c}^T) = \mathbf{0}$. There is no single objective function underlying O2PLS as a stepwise approach is used. In a first step, the shared variation is extracted from each data block as follows: Let $\mathbf{USV}^T$ be the singular value decomposition of the between-block covariance matrix $\mathbf{X}_2^T\mathbf{X}_1$. Then $\mathbf{E}_1 = \mathbf{X}_1 - \mathbf{X}_1\mathbf{VV}^T$ and $\mathbf{E}_2 = \mathbf{X}_2 - \mathbf{X}_2\mathbf{UU}^T$ could be considered some kind of residual matrices from which all common variation has been extracted. In a second step, the distinctive components of $\mathbf{X}_1$ are found one by one by maximizing, for $\mathbf{w}_{1,s}$, $\mathbf{w}_{1,s}^T\mathbf{E}_1^T\mathbf{X}_1\mathbf{VV}^T\mathbf{X}_1^T\mathbf{E}_1\mathbf{w}_{1,s}$ and, for $\mathbf{X}_2$, by $\max(\mathbf{w}_{2,s})$ $\mathbf{w}_{2,s}^T\mathbf{E}_2^T\mathbf{X}_2\mathbf{UU}^T\mathbf{X}_2^T\mathbf{E}_2\mathbf{w}_{2,s}$. The first distinctive component of $\mathbf{X}_1$ is $\mathbf{t}_{1,s} = \mathbf{E}_1\mathbf{w}_{1,s} = \mathbf{X}_1\mathbf{w}_{1,s}$ and of $\mathbf{X}_2$ is $\mathbf{t}_{2,s} = \mathbf{E}_2\mathbf{w}_{2,s} = \mathbf{X}_2\mathbf{w}_{2,s}$. The next distinctive components are found by repeating this operation after extraction of the previous distinctive components; for example, to extract the second distinctive components for the two data blocks, in the calculations (starting with the calculation of $\mathbf{E}_1$ and $\mathbf{E}_2$), $\mathbf{X}_1$ is replaced by $\mathbf{X}_1 - \mathbf{t}_{1,s}\mathbf{p}_{1,s}^T$ and $\mathbf{X}_2$ by $\mathbf{X}_2 - \mathbf{t}_{2,s}\mathbf{p}_{2,s}^T$. Denote the matrices obtained after the extraction of the final distinctive components by $\mathbf{X}_{res1}$ and $\mathbf{X}_{res2}$. In a final step, common components $\mathbf{T}_{1,c}$ for the first data block and $\mathbf{T}_{2,c}$ for the second data block are extracted one by one from the residual matrices $\mathbf{X}_{res1}$ and $\mathbf{X}_{res2}$ by relying on the MAXDIFF criterion [7]:

$$\max(\mathbf{w}_1, \mathbf{w}_2)\ \mathrm{tr}\left(\mathbf{w}_1{}^T\mathbf{X}_{res1}{}^T\mathbf{X}_{res2}\mathbf{w}_2\right) \text{ such that } \mathbf{W}_k{}^T\mathbf{W}_k = \mathbf{I} \quad \text{for all } k, \tag{18}$$

with $\mathbf{t}_{1,c} = \mathbf{X}_{res1}\mathbf{w}_1$ and $\mathbf{t}_{2,c} = \mathbf{X}_{res2}\mathbf{w}_2$.

The MAXDIFF criterion, unlike CCA (see expression (16)) also includes the variation of the data blocks; this means that the O2PLS common components account for the variation of the data block, but also that data blocks with a larger sum of squares have more weight in the analysis. To give equal weight to the data blocks, they may be scaled to equal sum of squares prior to the O2PLS analysis. In follow-up papers O2PLS is extended to the case of more than two linked data blocks [11] and to the case were multiset data can be linked object- and/or variable-wise [29]. In the latter paper, the simple case of variable wise data is treated by applying O2PLS to the transposed data. A closely related method, is the LS-ParPLS method proposed in a regression context by [19]: For multiblock data the method first extracts common components using canonical correlation analysis and subsequently unique components are extracted from each deflated data block.

3.6.2. Application to the illustrative data

3.6.2.1. Objectwise linked data. The application of O2PLS to the transcriptomics data obtained three and seven days after vaccination with TIV, is not trivial because the algorithm relies on the SVD of the matrix of cross-products between the variables (this is a 54,715 by 54,715 matrix) and on the SVD of an even larger matrix to optimize the MAXDIFF criterion. An efficient solution with limited memory requirements was implemented by relying on results and properties of the product SVD [30].

The proportion of VAF by four common components and one distinctive component for each data block is shown in Table 1. Note that the components C1–C5 represent five pairs of components that are orthogonal within blocks but, for the distinctive components, are also (almost) orthogonal within the pairs. The fact that the specific components of a pair are not the same but block-specific explains why C5 is simultaneously highlighted in yellow and pink. The differences in VAF between

[1] At the time of submitting the revised manuscript, the kernel extension to RGCCA was not yet submitted to CRAN but kindly made available by Arthur Tenenhaus.

Please cite this article as: K. Van Deun, et al., Identifying common and distinctive processes underlying multiset data, Chemometrics and Intelligent Laboratory Systems (2013), http://dx.doi.org/10.1016/j.chemolab.2013.07.005

data blocks are similar in size for the common and distinctive components. GSEA was applied with the ranking of the genes based on the $\mathbf{P}_{k,c}$ and $\mathbf{P}_{k,s}$ matrices in Eq. (17) for the first two components. The sets found for the first component consist of genes that are upregulated in T and B cells and in subjects vaccinated with LAIV or TIV flu vaccines. For the second component sets of genes are found that are upregulated in samples treated with TIV, LAIV flu vaccine, or LA yellow fever vaccine.

3.6.2.2. Variable-wise linked data. For ease of comparison with DISCO-SCA and the adapted GSVD, O2PLS was applied to the linked TIV- and LAIV-data blocks with three common components and three distinctive components. The results are summarized in Tables 3 and 4, under the header O2PLS. In Table 3 the VAF by each of the components in each data block and in total is shown; because the components are orthogonal within blocks, the total VAF equals the sum of VAF per component. Compared to (DISCO-)SCA and the adapted GSVD, the O2PLS components account for more variation in total. This is because the latter methods constrain the loadings to be the same for both data blocks while in 02PLS a set of loadings is estimated for each data block. Observe that the VAF by a specific component is considerably less than by a common component. Table 4 shows the correlations of the components with the antibody titers and the total R^2 when regressing the titers on the six components; these numbers were obtained for each data block separately. Within blocks the component scores are not orthogonal (though the loadings are) and thus the R^2 is not equal to the sum of squared correlations. Compared to (DISCO-)SCA and the adapted GSVD, the O2PLS components account for less variation in the titers for the subjects vaccinated with LAIV. The annotation was applied to a ranking of the genes based on the $\mathbf{T}_{k,c}$ and $\mathbf{T}_{k,s}$ matrices in Eq. (17). For the third component, with the ranking based on the scores associated to the TIV data, sets of genes are obtained that are upregulated in B cells, CD4 and CD8 T cells, and downregulated in subjects that acquired a viral or bacterial infection or that were vaccinated against yellow fever. When using the scores obtained for LAIV, sets of genes are found that are downregulated in B cells, CD4 and CD8 T cells and upregulated in subjects vaccinated against yellow fever. For the sixth component only a ranking of the scores associated to the TIV data block yield gene sets passing the threshold of a normalized enrichment score greater than 2. The sets found for TIV consisted of genes downregulated in B cells and upregulated in subjects that acquired a viral or bacterial infection. For all four rankings, no sets were found associated to TIV or LAIV vaccine.

3.7. Comparison of the results between all methods

For each of the methods, three types of performance were evaluated: the goodness of fit of the model to the data in terms of the proportion of variance accounted for, the predictive performance with respect to immunogenicity quantified by the plasma hemagglutination–inhibition antibody titers, and the biological content revealed in an enrichment analysis of the genes ranked by their component loadings or weights.

First, the proportion of VAF jointly by all five or six components is very similar between all methods except for O2PLS in the case of variable-wise linked data (compare the total VAF on the last line of Table 3). In that case, O2PLS accounts for approximately 5% more variation in each of the data blocks with a large contribution by the common components (C1–C3). O2PLS is less constrained than SCA, DISCO-SCA, and the adapted GSVD because the latter methods are subject to the same loadings for both TIV and LAIV data block while the O2PLS components may be very different.

The prediction of the antibody titers is best for the SCA-based methods: SCA, DISCO-SCA and the adapted GSVD have the highest R^2, see the last lines of Tables 2 and 4. These methods all yield the same R^2 as their components span the same low-dimensional space. Interestingly, in case of the objectwise linked data (Table 2) the Day 3 associated CCA and O2PLS components have more predictive power than the Day 7 data. In case of the variable-wise linked data (Table 4) prediction is better for the TIV group of vaccinees than for the group of LAIV vaccinees; for the latter group there were many low responders and thus less variation in the antibody titers.

On the level of interpretation of the objectwise linked data, SCA, DISCO-SCA, ECO-POWER, and regularized CCA find back the gene sets associated to the data from [1] and indicating differential expression in subjects vaccinated with TIV in particular and influenza vaccines in general. From an immunological point of view, common components are to be expected. The inclination of the adapted GSVD to yield distinctive components is less attractive in such a setting and a similar remark holds for O2PLS. For the variable-wise linked data, finding an early signature of the vaccine response that is predictive for vaccine efficacy is important and hence the focus is on differentiating more from less efficient vaccines and finding their differential mode of action. SCA is less suitable in this respect as it does not explicitly account for distinctive components. The results obtained with DISCO-SCA and the adapted GSVD were highly similar and resulted in gene sets associated to the differential mode of action of the TIV and LAIV vaccines as well as sets referring to a common mode of action. Although the gene sets in GSEA referring to TIV and LAIV come from the same data analyzed here, this is not a trivial result given that our analyses are based on data corrected for baseline per block while the sets in GSEA are based on differences in mean expression between TIV (or LAIV) post vaccination and baseline. Most interesting, also sets were recovered from another study, on yellow fever vaccine YF-$_{17}$D which is a live attenuated and very successful vaccine.

4. Discussion

So far, we gave a formal description of six methods for finding common and distinctive processes in linked data and we applied them to an empirical data set. Here, we will draw a basic comparison of the methods in question. A summary of this comparison is given in Table 5, the rows of which refer to the six methods (SCA, DISCO-SCA, adapted GSVD, ECO-POWER, CCA, and O2PLS) and the columns to five different characteristics.

A first characteristic is the applicability of the methods to linked objectwise and/or variable-wise data: All methods are applicable to multiset data that are linked objectwise; furthermore, SCA, DISCO-SCA, the adapted GSVD and O2PLS are also applicable to variable-wise linked data. For O2PLS an extension was proposed, called bi-modal OnPLS [29] that is applicable to multiblock data containing a mix of object- and variable-wise linked data. The natural extension of SCA and DISCO-SCA to such mixed cases seems to be one that simultaneously uses the same quantification of the component scores and loadings for blocks that are linked object- and variable-wise respectively. The paper was limited to the case of two data blocks; except for the adapted GSVD all methods can also be applied with more than two data blocks. However, in the case of a large number of blocks the option to partition the blocks in groups of blocks with the same underlying specific component(s) may be interesting; for this purpose, a clusterwise SCA approach that yields common (shared between all components) and group-specific components has recently been proposed [31].

A second characteristic pertains to whether the modeling involves a single quantification for the linking mode. This is, for example, the case for SCA where the same matrix of component scores is used for both data blocks with objectwise linked data, and the same matrix of loadings for variable-wise linked data. Also DISCO-SCA, the adapted GSVD, and ECO-POWER imply a single quantification of the linking mode for both data blocks. The canonical variates of CCA differ between the data blocks and also O2PLS results in a different quantification of the linking mode for the two data blocks. The CCA and the O2PLS common variates have respectively maximal correlation or covariance between blocks and the O2PLS specific components have low correlation. The advantage of a single quantification of the linking mode is that this has a clearer meaning than, for example, two sets of quantifications that

Please cite this article as: K. Van Deun, et al., Identifying common and distinctive processes underlying multiset data, Chemometrics and Intelligent Laboratory Systems (2013), http://dx.doi.org/10.1016/j.chemolab.2013.07.005

Table 5
Summary of a comparison of the six methods for identifying common and distinctive mechanisms (SCA, DISCO-SCA, Adapted GSVD, ECO-POWER, CCA, and O2PLS) in terms of the scope of their applicability, the type of quantification used for the linking mode, the criterion used for estimation, whether within-block VAF is taken into account, whether the components are orthogonal within blocks either on the level of scores or loadings, and the type of formalization for common and distinctive mechanisms. In the cells, an 'x' indicates that the characteristic applies to the method, 'NA' means that it is not applicable.

Method	Applicability		Single quantification linking mode	Global criterion	Takes within-block VAF into account	Orthogonality within blocks		Formalization	
	Linked objectwise	Linked variable-wise				Linked objectwise	Linked variable-wise	Common	Distinctive
SCA	x	x	x	x	x	Scores	Loadings	NA	NA
DISCO-SCA	x	x	x	x	x	Scores	NA	Residual status	Zero scores/loadings
Ad.GSVD	x	x	x	NA	x	Loadings	Scores	Residual status	Maximal relative VAF
ECO-POWER	x	NA	x	x	x	Scores	NA	Product %VAF	NA
CCA	x	NA	NA	x	x	Scores	NA	Maximal correlation	NA
O2PLS	x	x	NA	NA	x	Scores	Loadings	Maximal covariance	Orthogonal to common base

correlate highly: In the latter case there is no certainty that the two sets of quantifications represent the same structural differences between the elements of the linking mode. A single quantification in case of distinctive components implies that these are truly *distinctive* components in the sense that they underlie one of the data blocks and at the same time do not underlie the other data block. This is very different from the O2PLS specific components that are close to orthogonal within a pair. Another aspect related to the interpretation of such components is that the methods that rely on a single quantification for the linked mode yield components that lie in the space spanned by all data blocks simultaneously. In case the subspaces spanned by the blocks are orthogonal, this may yield common components that lie in between these subspaces and with considerable VAF in each block, for example when using ECO-POWER to find common components. Methods that rely on a quantification per block would yield orthogonal components. Whether such components are 'common' is not a trivial discussion; see also [19] on a discussion about the interpretation of components based on a linear combination of the variables of the two data blocks.

A third characteristic is the use of a global optimization criterion to model the data. The advantage of such a criterion is that all available information is used at once to estimate the different submodels pertaining to the different data blocks. SCA, DISCO-SCA, ECO-POWER, and CCA all rely on such a global optimization criterion from which all estimates (for all components in all data blocks) are obtained at once. To the best of our knowledge there is no such a global criterion for the adapted GSVD, hence the 'NA' in Table 5; yet, the GSVD procedure does yield a simultaneous estimation of all model matrices at once and results in a least-squares approximation to the concatenated data. O2PLS does not rely on a global criterion but works in a sequential way: First the common base is removed from the data, then the distinctive components are extracted and removed from the original data, and, finally, the common components are extracted. In the different steps of this procedure different criteria are being used.

A fourth characteristic is whether accounting for variance within the data blocks is included in the optimization. This is the case for all methods except for ordinary canonical correlation analysis (unlike generalizations of it: see [7]). SCA, DISCO-SCA, and the adapted GSVD are least-squares methods and their sum of block-specific VAF is maximal. ECO-POWER also relies on an optimization criterion that maximizes the block-specific VAF but combines the block-specific contributions in terms of their product instead of their sum.

The orthogonality of the components within blocks is a fifth characteristic on which the methods can be compared. In Table 5 the orthogonality within the data blocks is considered and, if applicable, it is specified whether this orthogonality stems from the scores or loadings. For objectwise linked data, orthogonality always applies and, except for the adapted GSVD, stems from the matrix of component scores (in case of the adapted GSVD, the loadings of each block are orthogonal). Hence, in a linear regression context, the components contribute in a pure way to the variance of an external outcome variable for SCA, DISCO-SCA, ECO-POWER, and O2PLS. For variable-wise linked data orthogonality does not apply to DISCO-SCA because both the rotated loadings and the component scores at the level of the block are non-orthogonal (note, however, that the component scores are orthogonal over blocks). As a consequence of this, the DISCO-SCA components do not contribute in a pure way to the variance accounted for jointly by all components in a block. SCA and O2PLS have orthogonal loadings with variable-wise linked data while the adapted GSVD has orthogonal component scores at the level of the blocks. That the distinctive components of a data block are orthogonal to those of the other data block and also to the common components seems to be a natural requirement given the concept of 'distinctiveness'. Whether in case of variable-wise linked data the orthogonality should be at the level of the concatenated data or within the data blocks is subject to discussion (see [13,31]).

The sixth characteristic pertains to the formalization of common and distinctive components. SCA extracts components without explicit focus on common or distinctive; hence, the SCA components are a mix of common and distinctive sources of variation. Note, however, that SCA favors common components because the sum of the block-specific VAFs of a common component will usually be higher than the sum of the block-specific VAFs of a distinctive component. DISCO-SCA and the adapted GSVD focus primarily on finding distinctive components; DISCO-SCA does so by rotating the SCA solution to a target with zero-scores or loadings for the distinctive components, and the adapted GSVD by transforming the SCA solution to one with maximal contribution to the VAF in one data block relative to the other. This means that distinctive is formalized as a component that is only underlying one data block while at the same time being explicitly absent in the other data block. The common components of DISCO-SCA and the adapted GSVD have a residual status in the sense that they are not targeted directly in the optimization. Otherwise, there may be a bias of these methods to common components as they are derivatives of the SCA solution. ECO-POWER and CCA focus on common components. ECO-POWER does so by maximizing the product of the block-specific VAF by a component, and CCA by maximizing the correlation between the canonical variates. O2PLS relies on a formalization of both common and distinctive components: The former are subject to the MAXDIFF criterion, which takes both the within-block VAF and the correlation between blocks into account, whereas the latter are constrained to be orthogonal to the joint space of the two data blocks while having simultaneously maximal overlap with the so-called predictive space of the data block for which they are distinctive.

To conclude, we presented six methods for finding common and distinctive processes in multiset data that are either linked by the objects or by the variables. Their formal description was given and all six methods were applied to gene expression data and the results were used to predict vaccine efficacy and were interpreted by an annotation tool. We also gave a basic comparison of the methods. SCA, DISCO-SCA, and O2PLS are the most flexible methods because they can be used both with objectwise and variable-wise linked data and also with

Please cite this article as: K. Van Deun, et al., Identifying common and distinctive processes underlying multiset data, Chemometrics and Intelligent Laboratory Systems (2013), http://dx.doi.org/10.1016/j.chemolab.2013.07.005

more than two data blocks. Among these three methods, DISCO-SCA may be preferred over the other two because it disentangles common and distinctive sources of variation (unlike SCA) and uses a single quantification for the linked mode (unlike O2PLS).

Acknowledgments

This research was supported by the Belgian Federal Science Policy (IAP P7/06). We would like to thank two anonymous reviewers for their useful comments and Arthur Tenenhaus for his kind assistance with the RGCCA package and for providing an early version of the updated RGCCA package.

References

[1] H.I. Nakaya, J. Wrammert, E.K. Lee, L. Racioppi, S. Marie-Kunze, W.N. Haining, A.R. Means, S.P. Kasturi, N. Khan, G.-M. Li, M. McCausland, V. Kanchan, K.E. Kokko, S. Li, R. Elbein, A.K. Mehta, A. Aderem, K. Subbarao, R. Ahmed, B. Pulendran, Systems biology of vaccination for seasonal influenza in humans, Nature Immunology 12 (2011) 786–795.
[2] R. Tauler, A.K. Smilde, B. Kowalski, Selectivity, local rank, three-way data analysis and ambiguity in multivariate curve resolution, Journal of Chemometrics 9 (1995) 31–58.
[3] A.K. Smilde, M.J. van der Werf, S. Bijlsma, B.F. van der Werff-van der Vat, R.H. Jellema, Fusion of mass spectrometry-based metabolomics data, Analytical Chemistry 77 (2005) 6729–6736.
[4] S. Bergmann, J. Ihmels, N. Barkai, Similarities and differences in genome-wide expression data of six organisms, PLoS Biology 2 (2004) e9.
[5] J. Jaumot, R. Tauler, MCR-BANDS: a user friendly MATLAB program for the evaluation of rotation ambiguities in Mulitvariate Curve Resolution, Chemometrics and Intelligent Laboratory Systems 103 (2010) 96–107.
[6] P. Tamayo, D. Scanfeld, B.L. Ebert, M.A. Gillette, C.W.M. Roberts, et al., Metagene projection for cross-platform, cross-species characterization of global transcriptional states, Proceedings of the National Academy of Sciences of the United States of America 104 (2007) 5959–5964.
[7] M. Hanafi, H.A.L. Kiers, Analysis of K sets of data, with differential emphasis on agreement between and within sets, Computational Statistics and Data Analysis 51 (2006) 1491–1508.
[8] H.A.L. Kiers, J.M.F. ten Berge, Hierarchical relations between methods for simultaneous component analysis and a technique for rotation to a simple simultaneous structure, British Journal of Mathematical and Statistical Psychology 47 (1994) 109–126.
[9] K. Van Deun, A.K. Smilde, M.J. van der Werf, H.A.L. Kiers, I. Van Mechelen, A structured overview of simultaneous component based data integration, BMC Bioinformatics 10 (2009) 246.
[10] M. Schouteden, K. Van Deun, I. Van Mechelen, ECO-POWER: a novel method to reveal common mechanisms underlying linked data, in: A. Colubi, K. Fokianos, E.J. Kontoghiorghes (Eds.), Proceedings of COMPSTAT'2012. 20th International Conference on Computational Statistics, Physica-Verlag, Heidelberg, 2012, pp. 757–768.
[11] T. Löfstedt, J. Trygg, OnPLS — a novel multiblock method for the modelling of predictive and orthogonal variation, Journal of Chemometrics 25 (2010) 441–455.
[12] O. Alter, P.O. Brown, D. Botstein, Generalized singular value decomposition for comparative analysis of genome-scale expression data sets of two different organisms, Proceedings of the National Academy of Sciences of the United States of America 100 (2003) 3351–3356.
[13] M. Schouteden, K. Van Deun, S. Pattyn, I. Van Mechelen, SCA with rotation to distinguish common and distinctive information in linked data, Behavior Research Methods (2013), http://dx.doi.org/10.3758/s13428-012-0295-9, (in press).
[14] K. Van Deun, I. Van Mechelen, L. Thorrez, M. Schouteden, B. De Moor, M.J. van der Werf, L. De Lathauwer, A.K. Smilde, H.A.L. Kiers, DISCO-SCA and properly applied GSVD as swinging methods to find common and distinctive processes, PLoS One 7 (e37840) (2012) 1–13.
[15] B.M. Bolstad, R.A. Irizarry, M. Astrand, T.P. Speed, A comparison of normalization methods for high density oligonucleotide array data based on bias and variance, Bioinformatics 19 (2) (2003) 185–193.
[16] A. Subramanian, P. Tamayo, V.K. Mootha, S. Mukherjee, B.L. Ebert, M.A. Gillette, A. Paulovich, S.L. Pomeroy, T.R. Golub, E.S. Lander, J.P. Mesirov, Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles, PNAS 102 (2005) 15545–15550.
[17] H.A.L. Kiers, Towards a standardized notation and terminology in multiway analysis, Journal of Chemometrics 14 (2000) 105–122.
[18] M.E. Timmerman, H.A.L. Kiers, Four simultaneous component models for the analysis of multivariate time series from more than one subject to model intraindividual and interindividual differences, Psychometrika 68 (2003) 105–121.
[19] I. Måge, B.-H. Mevik, T. Næs, Regression models with process variables and parallel blocks of raw material measurements, Journal of Chemometrics 22 (2008) 443–456.
[20] S. Friedland, A new approach to generalized singular value decomposition, SIAM Journal on Matrix Analysis and Applications 27 (2005) 434–444.
[21] C.C. Paige, M.A. Saunders, Towards a generalized singular value decomposition, SIAM Journal on Numerical Analysis 18 (1981) 398–405.
[22] J.M.F. ten Berge, Least Squares Optimization in Multivariate Analysis, DSWO Press, Leiden, 1993.
[23] H.A.L. Kiers, A.K. Smilde, A comparison of various methods for Multivariate Regression with highly collinear variables, Statistical Methods and Applications 16 (2007) 193–228.
[24] T. Dahl, T. Næs, A bridge between Tucker-1 and Carroll's generalized canonical analysis, Computational Statistics & Data Analysis 50 (2006) 3086–3098.
[25] A. Tenenhaus, M. Tenenhaus, Regularized generalized canonical correlation analysis, Psychometrika 76 (2011) 257–284.
[26] A. Tenenhaus, RGCCA Package. Downloaded from http://cran.r-project.org/web/packages/RGCCA/index.html 2012.
[27] A. Tenenhaus, et al., Kernel generalized canonical correlation analysis, Journal of Machine Learning Research (2013), (submitted for publication).
[28] J. Trygg, O2-PLS for qualitative and quantitative analysis in multivariate calibration, Journal of Chemometrics 16 (2002) 283–293.
[29] T. Löfstedt, L. Eriksson, G. Wormbs, J. Trygg, Bi-modal OnPLS, Journal of Chemometrics 26 (6) (2012) 236–245.
[30] B. de Moor, On the structure and geometry of the product singular value decomposition, Linear Algebra and its Applications 168 (1992) 95–136.
[31] K. De Roover, M.E. Timmerman, B. Mesquita, E. Ceulemans, Common and cluster-specific simultaneous component analysis, PLoS One 8 (5) (2013) e62280.

Please cite this article as: K. Van Deun, et al., Identifying common and distinctive processes underlying multiset data, Chemometrics and Intelligent Laboratory Systems (2013), http://dx.doi.org/10.1016/j.chemolab.2013.07.005